DESIGNING A LIST FOR WORD SQUARES

LEONARD GORDON
Tucson, Arizona

In "Bottoms Up!" in the February 1993 Word Ways, Eric Albert asserts that searching for word squares with or without a computer is faster if the search is made bottom-up. In "Which Way to the Square?" in the May Word Ways, Richard Sabey contradicts this assertion; he finds that for 8x8 squares from the Official Scrabble Players Dictionary (OSPD) top-down is considerably faster.

At one point when refuting Ted Clarke's opinion that top-down is faster, Eric used the words "well-written program and well-designed database". What did he mean by that? He was probably correct in saying that Ted Clarke's view was based on a poorly written computer program. But, programs used by others are well written. I doubt if any of them would contradict Richard Sabey. But this is only half of Eric Albert's point. Whether he realized it or not, neither the OSPD nor Webster's Second Unabridged (Web 2) produce a well-designed database. They are not poorly designed; they are simply not designed for this.

In "Word-Square Support: Part 2" in the November 1993 Word Ways, I showed that a database (list) can be improved by increasing the proportion of words starting with a vowel. I worked with 6-letter words, but my plan was to gain knowledge for building efficient lists of longer words. I now continue my study, this time using 8-letter words--and, while at it, examine the top-down vs. bottom-up argument.

My master list contains 51,077 8-letter words. 47,607 are lower case from Web 2 and elsewhere; 26,440 of them are from OSPD. The 3,370 capitalized words are all from Web 2. Using a similar but all lower case list, Richard Sabey got 315 squares from 45,594 words (see "Some More Quality Eight-Squares" in the August 1995 Word Ways). His May study got two squares from the OSPD. Taking the usual figure of merit,

$$Q = 45{,}594/315^* = 22{,}200$$
$$Q = 26{,}444/2^* = 24{,}200$$

where * denotes "take the eighth root of". Table 1 at the end of this article presents statistics for my list. All numbers are rounded percentages. Within a box, the three columns refer to three stocks within the list. Stocks A, B, and C are, respectively, OSPD, lower case words not in OSPD, and capitalized words. Table 1 shows a lot of unbalance in the stock. Note that in stocks A and B, only 2 and 4 per cent of the words end with "a", whereas "a" constitutes 8 per cent of the letters in positions 1-7. This is an unbalance. For "a" in first position, the lists

are fairly well balanced. With all three stocks, only 4 per cent of the words begin with "e" compared to an average of about 11 per cent of the letters in matching positions. This is unbalanced. "e" in eighth position is fairly well balanced. The worst unbalance is with "s" in eighth position. Thirty-four per cent of 8-letter OSPD words end in "s". I will not enumerate all the other balances and unbalances.

Let's see what happens if we attempt to balance our list by judiciously selecting 20,000 or so words from the 50,000. Two tests were made. The selection process for the first took all words ending "a" or "i" and all words beginning "e" "i" "n" or "o". It then accepted all words with a lot of "s" "d" or "y" in positions 1-7, excluding those containing "o" or "u" or other letters that seldom occur in the eighth position. Finally, the list was filled from stock A, skipping every third word and skipping every sixth word that ended in "s". This list produced 11 squares from 22,476 words. Selection for the second test did the same thing except that it also accepted words from the C stock that began with "a", and it skipped more words from stock A ending in "s". It also found 11 squares. Figures of merit are

Q = 22,476/11* = 16,700
Q = 19,287/11* = 14,291

Long"s support value ("Mathematics of Square Construction", February 1993 Word Ways) for the eight-square is 15,678.

Table 2 summarizes letter distributions for my master list and the two test lists. Within each box, the first column gives the percentage occurrences of letters in 1st position, the second column is for the 8th position, and the third column gives the percentage at which the letter occurs in the stock. Note that there is still a lot of unbalance in the test lists, but I did make the lists more productive than the OSPD. We can probably get lower support values by further shortening the list, but this gets into a region where squares are infrequent and statistics consequently not very reliable. It would be better to go outside Web 2 to get "good" words. Let's leave that for later.

I use a 33 mhz MS/DOS 386-DX computer, and program in Microsoft QuickBasic. Running uncompiled, test #2 took 15.14 hours to exhaust the search. With a reversed word-list (i.e., searching bottom-up) the run took 16.30 hours. Richard Sabey found that bottom-up took 61 per cent longer. His finding is probably exaggerated due to the large unbalances in the OSPD, but in any case I see no point in searching bottom-up with a computer unless we are using heuristics. With regard to heuristics, the method described by Chris Long in the May 1993 Word Ways to select bottom words seems like a good idea. Augment your list as well as possible using the ideas in this article, and then determine your own formula for word quality. With this scheme, searching bottom-up may well be faster than top-down.

The results here are not very encouraging for finding 10x10 squares. Long's 10x10 support value is 248,000 words. I had hoped that studies on 8x8 squares would project to something smaller. Although my selection process resulted in lists far better than Web 2 or the OSPD, it was mainly because the latter was so poor to start with. My best was little better than Long's 8x8 support value. Nevertheless, let's not give up hope. I suggest that Word Ways readers who may be tired of -gry words find some good words for us computer guys.

TABLE 1

	BEGINNING LETTER			LETTERS 2-8			FINAL LETTER			LETTERS 1-7		
	A	B	C	A	B	C	A	B	C	A	B	C
a	6	6	11	7	8	14	2	4	23	8	8	12
b	6	7	6	1	1	1	0	0	0	2	2	2
c	9	8	10	3	3	3	2	3	4	4	3	4
d	6	5	4	3	3	2	11	11	2	3	2	3
e	4	4	4	11	11	8	11	17	15	10	9	7
f	5	4	2	1	1	0	0	0	0	1	1	1
g	4	3	5	3	2	1	8	3	1	2	2	2
h	3	3	4	2	2	2	1	2	2	2	2	3
i	3	2	2	8	8	10	0	1	3	8	8	10
j	1	1	2	0	0	0	0	0	0	0	0	0
k	1	1	2	1	1	1	1	2	0	1	1	1
l	3	3	4	5	5	4	3	5	2	5	5	5
m	5	5	8	2	3	3	2	4	6	3	3	3
n	2	2	2	6	6	7	4	7	18	6	6	5
o	3	4	2	6	7	6	0	1	2	6	7	6
p	8	8	8	2	2	2	0	1	0	3	3	3
q	1	0	1	0	0	0	0	0	0	0	0	0
r	6	4	3	7	6	6	7	6	2	6	6	6
s	12	11	9	8	6	5	34	19	14	5	5	5
t	5	6	7	5	5	4	6	7	3	5	5	5
u	3	6	1	3	3	4	0	0	1	3	4	4
v	2	2	2	1	1	1	0	0	0	1	1	1
w	3	3	1	1	1	0	0	0	0	1	1	0
x	0	0	0	0	0	0	0	0	0	0	0	0
y	0	0	1	2	2	2	6	6	2	1	1	1
z	0	0	0	0	0	0	0	0	0	0	0	0

TABLE 2

	ENTIRE LIST 51,077			TEST #1 22,476			TEST #2 19,287		
a	6	4	8	5	9	8	6	11	9
b	6	0	3	5	0	2	5	0	2
c	9	3	4	8	2	4	7	3	4
d	5	10	4	5	10	4	4	8	4
e	4	14	11	9	12	12	10	14	12
f	4	0	2	4	0	1	3	0	1
g	3	5	3	3	7	3	3	7	3
h	4	2	3	3	1	2	3	2	2
i	2	1	8	5	2	8	6	2	9
j	1	0	0	1	0	0	1	0	0
k	1	1	1	1	1	1	1	1	1
l	3	4	5	3	3	5	2	4	5
m	5	3	3	5	2	3	4	3	3
n	2	6	6	4	4	6	5	5	7
o	3	1	6	8	1	7	9	1	7
p	8	1	3	6	0	3	6	1	3
q	1	0	0	0	0	0	0	0	0
r	5	6	7	4	7	7	4	7	7
s	12	26	9	10	25	8	10	20	8
t	6	6	6	5	6	6	4	7	6
u	4	0	4	2	0	3	2	0	3
v	2	0	1	1	0	1	1	0	1
w	2	0	1	2	0	1	2	0	1
x	0	0	0	0	0	0	0	0	0
y	0	6	2	0	6	1	0	5	1
z	0	0	0	0	0	0	0	0	0